Mapping of Genomic Structural Variations in Major Birth Defects

Study protocol

Version: V 1.0 Date: February 19,2026

Project Background

Undertaking etiological research on birth defects is of paramount importance for their prevention and control. Beyond environmental factors, genetic variations represent a significant contributor to the occurrence of defective offspring among healthy parents. The Whole Exome Sequencing (WES) technology, which is based on second - generation short - read sequencing, exhibits limited comprehensive coverage of genetic variation types, thereby imposing constraints on genetic diagnosis. Structural Variations (SVs), characterized by high mutation rates, substantial degrees of variation, and strong pathogenicity, have emerged as a research focus in the field of life sciences in recent years. Presently, the commonly employed genetic testing methods in clinical practice possess insufficient resolution for long repeat sequence regions and SVs, which may give rise to missed diagnoses, thus necessitating the introduction of novel technical approaches. Long - read sequencing represents the developmental direction for genomic variation (especially SV) analysis.

In addition to technological innovation, genetic etiological research on major birth defects in the Chinese population still needs to surmount two technical challenges. One of these challenges is the requirement to construct a genomic reference map. There are notable disparities in genomic structure and sequence among different populations, yet the current human genomic reference map is based on Caucasian samples. Applying this map to identify genomic SVs in the Chinese population may result in severe misinterpretations. Therefore, it is imperative to establish a pan - genomic reference map for the Chinese population.

This study will concentrate on challenging cases with unclear genetic diagnoses, conducting long - read DNA sequencing data analysis of birth defect cases and family samples. It will prioritize the extraction and identification of individual - specific genomic features, develop detection algorithms for all types of structural variations (SVs), including complex SVs, and establish a pan - genomic reference map exclusive to the Chinese population. This will facilitate the identification of pathogenic SVs in birth defect cases and family samples of the Chinese population, as well as the mapping of fine SV profiles for major birth defects in the Chinese population**.**

Research Objectives and Significance

In cases of prenatal genetic diagnosis with ambiguity and complexity, this project intends to carry out long - read DNA sequencing data analysis on birth defect cases and family samples. The emphasis lies on the extraction and identification of individual - specific genomic characteristics, along with the development of detection algorithms for all types of structural variations (SVs), including complex SVs. It will establish a pan - genomic reference map specific to the Chinese population, which can support the identification of pathogenic SVs in the birth defect cases and family samples of the Chinese population, and map the detailed SV spectrum of major birth defects in the Chinese population. Additionally, the project will conduct in - depth analysis of the genetic and pathogenic roles of different types of SVs in birth defects, offering a theoretical foundation for promoting the early warning, intervention, and prevention of major birth defects in China.

Research Content and Protocol

Based on the long-read DNA sequencing of cases carried out in Project Topics 1 and 3, this study incorporates case and control samples, along with their long-read sequencing data. The PacBio Revio platform was chosen for long-read DNA sequencing, and standard whole-genome long-read sequencing analysis was conducted on 50 cases. Building on the first-phase China population pan-genome reference map that has already been constructed, 50 representative control samples were added for long-read sequencing to construct a new China population pan-genome reference map. The details are as follows:

(1) Extraction and encoding of genomic SV features in cases of birth defects, establishing an SV detection method based on feature encoding

Sequence visualization is used to eliminate background repetitive information from site alignment signals, thus enabling SV detection in complex genomic regions. A "compressed" image encoding method is explored, using the "stacking" technique to represent the abnormal sequence features between birth defect cases, parental controls, and population controls in a single image. Based on whole - genome alignment results, the genomic regions containing abnormal sequences in birth defect cases are identified. A local breakpoint - sensitive realignment method based on collinear segments is established to extract SV fragment characteristics from sequencing reads between birth defect cases, parental controls, and population controls. The study focuses on developing an isomorphic convolutional neural network framework capable of simultaneous target segmentation and classification, achieving both segmentation and classification in SV detection.

(2) Construction of a Chinese population pan - genome reference map based on long - read DNA sequencing data

Using third-generation whole-genome sequencing technology, DNA samples from multiple ethnic groups in China were sequenced, and visualization of pan-genome assembly was achieved. Ethnic-specific reference genomes were constructed to create pan-genome graphs, integrating DNA sequences from different populations. Genetic variations or sequences with differences were regarded as nodes, and adjacent sequences were connected by edges. This approach identified core and specific gene sequences in the Chinese population, thereby establishing a high-quality pan-genome reference map exclusive to China's population.

The study also focused on whole-genome SV mapping to support the precise analysis of rare or novel SVs in birth defects.

(3) Mapping the SV fine atlas of major birth defects in China

SV detection is carried out through two approaches: the linear genome and the pan - genome. The linear genome approach uses conventional linear detection methods to identify genetic variations. The pan - genome approach constructs a genome map by combining de novo assembled genomes with the universal human reference genome (GRCh38), along with cases of genetic variations and birth defects found in China. These two approaches mutually validate and complement each other, integrating the obtained SV results. Thresholds are set based on criteria such as the location of variations and sequence similarity, and redundant results are removed.

Methods, technical routes and feasibility analysis

Methods and Technical Approaches

1、This study aims to conduct structural variation (SV) analysis in complex cases of birth defects using third - generation sequencing platforms such as PacBio. The workflow encompasses the following steps: extraction of SV characteristic sequences, “compressed” structural variant feature encoding, SV detection, and performance evaluation.

The remaining samples from the enrolled cases underwent three - generation sequencing analysis of family members. Meanwhile, the SVision method developed by Xi'an Jiaotong University in the early stage of this project was utilized to analyze complex structural variations (CSVs), as detailed below:

- Generate denoised images: Collect abnormal long read pairs, create variant allele reference (VAR) and reference (REF) images, and obtain VAR - to - REF images by subtracting REF - to - REF images to reduce false positives introduced by repetitive sequences.

- tMOR procedure: Obtain single - variant images using a two - step image segmentation process, define regions around each breakpoint as segments of interest (SOI), and identify SOIs through a pre - trained CNN to form CSVs.
- The Grapher component employs a graph - based approach to represent various CSV structures. By detecting isomorphic graph merging of given CSV graph structures and their topological equivalence events, it generates the reference graph fragment assembly (rGFA) format for CSV graphs.
- Event confidence measurement: Cluster single - variant images that support the same event, integrate CNN prediction probabilities and the similarity between single - variant images to measure event confidence.

2. Further improve the mapping of the pan - genome atlas of the Chinese population

Based on the first - phase pan - genome reference map of the Chinese population that has already been constructed, 50 representative control samples were added. Using third - generation whole - genome sequencing technology, de novo sequencing was performed on multi - ethnic DNA samples from the Chinese population. The SVision analysis method was employed, with GRCh38 as the reference, to identify complex structural variations (CSV), achieve visualization of pan - genome assembly, and detect and record the core and specific gene sequences of the Chinese population. This effort aims to establish a high - quality pan - genome reference map exclusive to the Chinese population.

3. Mapping the SV Fine Atlas of Major Birth Defects in China

Using published methods, case sequencing data is aligned with the China Population Pan - Genome Reference Map. Based on the alignment results, the abnormal alignment regions are identified, and all abnormal alignment reads located in these regions are extracted. A multi - objective recognition framework based on equivariant convolutional neural networks is employed to detect various types of SVs.

SV detection is carried out via two approaches: the linear genome and the pan - genome.

① Linear genomic pathway: Conventional linear detection methods are utilized to identify genetic variations.

② Pan - genome approach: A genome map is constructed by combining de novo assembled genomes with the universal human reference genome (GRCh38) and genetic variant and birth defect case samples found in Chinese populations.

③ The results of the two methods can be mutually verified and supplemented.

SV fragment characteristics are extracted from sequencing reads between case -

parent controls and population controls. The SV results obtained from both approaches are integrated. Thresholds are set based on criteria such as mutation location and sequence similarity, and redundant results are removed.

On this basis, the SV fine map of major birth defects in the Chinese population is drawn.

Feasibility Analysis of the Task

The research team consists of Peking Union Medical College Hospital, Chinese PLA General Hospital, Fudan University, and Xi'an Jiaotong University. The members and institutions previously led the development of application guidelines for CMA technology in prenatal diagnosis, released the first China population pan - genome reference map, and developed a novel deep - learning - based SV analysis method (SVision), achieving precise identification of complex SVs. The solid foundation of the preliminary work provides strong technical support for the project, demonstrating its full feasibility.

Selection criteria, contraindications, and adverse reactions of the study subjects

1) Inclusion criteria:

• Single pregnancy with ultrasound findings indicating fetal structural abnormalities.

• Negative results for prenatal WES, karyotyping, CMA, etc.

• Alternatively, only one heterozygous pathogenic variant is detected in a suspected recessive genetic disorder, with no second suspected pathogenic variant identified

2) Exclusion criteria:

• Twin/multiple pregnancy

• No interventional prenatal diagnosis performed

• Refusing further testing

3) Adverse reactions: This study only utilized the remaining samples obtained for testing, with no adverse effects observed in patients.

Data statistics method

Statistical analysis was carried out by statisticians using SPSS 22.0 software. Measurement data were described as mean ± standard deviation ($\bar{x} \pm s$), median, maximum, minimum, and quartiles. Categorical data were expressed as percentages (%). All hypothesis tests were two - tailed, with $P < 0.05$ regarded as statistically significant.

Baseline data comparability was evaluated using a two - tailed statistical test at an alpha level of 5%. Categorical data comparisons were performed using the chi - square

test or Fisher's exact probability method, while measurement data comparisons were conducted with t - tests. Nonparametric variable comparisons were analyzed using the rank - sum test.

Statistical analysis planning was completed by professional statisticians. After all data entry and review were finalized, statisticians promptly completed the statistical analysis and issued a written statistical analysis report.

Potential risks of the project

1) Due to the limitations of current medical testing technologies, individual variations, and other known or unpredictable factors, the accuracy, sensitivity, and specificity of this test are restricted, with the potential for false - negative or false - positive results. All tests related to this study have certain limitations, so participants are fully informed prior to enrollment.

2) Given the current limitations in cognitive understanding, test results may show findings with unclear clinical significance or no positive detection. Therefore, participants must be fully informed and understand this.

3) Test results should be combined with genetic counseling.

4) The detection subject is limited to the specific indicators being examined, and this study cannot reflect the underlying diseases or risks present in the patient.

Data Preservation and Confidentiality

The medical information, specimen materials, and test data obtained from this research project will be strictly confidential.

1) When the test results are published in academic journals, no personally identifiable information will be disclosed.

2) During the project implementation period, the laboratory shall preserve the remaining samples for subsequent verification, and no unauthorized individuals shall access these specimens. Upon project completion, the remaining samples shall be stored, processed, or destroyed in accordance with clinical laboratory requirements. Test data shall be uniformly preserved as per hospital regulations, with measures implemented to ensure information security.

3) The human genetic resources involved in the project shall be registered with the Ministry of Science and Technology in accordance with the "Regulations on the Registration of Important Genetic Families and Specific Regional Human Genetic Resources" issued by the Ministry of Science and Technology.

Confidentiality measures for subject information

1) This test does not collect the subject's ID number or any other information that may disclose patient privacy.

2) The collected basic information and specimen data shall be kept strictly confidential. They are accessible only to authorized personnel for review and analysis, and no other individuals are permitted to obtain such materials.

3) Subsequent research findings shall not be used for any non - research purposes. In the case of publishing research articles, the information of participants will be strictly kept confidential.